A Creativity Survey of Unsupervised Graph Neural Network: Interactive Clustering and Embedding

XIA LIU

School of Computer and Information Engineering, Henan Normal University; lx15518903750@163.com

WEI LI

Institute of Computing Technology, Chinese Academy of Sciences; liwei@ict.ac.cn

Abstract: The rise and application of neural network has successfully promoted the research of pattern recognition and data mining. In recent years, graph neural network has attracted more and more attention. It has some applications in text classification, sequence annotation, neural machine translation, relation extraction, image classification and other fields. This review mainly integrates the existing research on semi-supervised or unsupervised graph neural network. The research work of this paper is mainly classified in three aspects, one is based on the classification of research questions, the other is based on the classification of research methods, and the third is based on the classification of measures .The main research problems are the low-dimensional representation of nodes in graphs and the over-smooth problem in the process of message transfer. The research methods mainly focus on the graph embedding algorithm, such as the graph embedding algorithm based on probability graph and the method based on deep learning. The measurement methods mainly focus on the accuracy and efficiency of the algorithm and model .Finally, this paper also puts forward the feasible future research direction, which provides reference for readers.

Keywords: Graph neural network, unsupervised learning, network embedding, node clustering

1 INTRODUCTION

Graph, also called network, as a common data structure, widely appears in People's Daily life, such as social network, World Wide Web and so on. The rise and application of neural network has successfully promoted the research of pattern recognition and data mining. Many machine learning tasks that once relied heavily on manual feature extraction, such as object detection machine translation and speech recognition, have been revolutionized by various end-to-end deep learning paradigms, such as convolutional Neural networks (CNN) long and short term memory (LSTM) and auto-encoders. Graph neural networks is a kind of link model, which relies on the information transfer between nodes in the Graph to capture the dependency relations in the Graph. Most graph neural networks (GNNS) are specifically designed for semi-supervised learning tasks (transduction and induction), in which supervised information (labeled nodes) plays an important role in enhancing the resolution. Unfortunately, supervised information cannot fundamentally solve this problem for two reasons: first, embedding in the network, unsupervised information cannot be obtained from unsupervised tasks such as node clustering and link prediction. Most existing unsupervised GNNS either reconstruct the original information (adjacency matrix and attribute matrix) or maximize mutual information to retain as much information as possible. Therefore, the super-smooth problem of unsupervised GNNS tends to be worse than that of semi-supervised GNNS. Second, while learning message aggregation from labeled nodes can alleviate smoothing problems, it can also cause serious over-fitting problems that can severely impact performance.

This paper on the existing research literature about the diagram of the neural network are classified, based on the literature reading and study, first of all, we analyzed the various literature research, found that part of the literature is to more complex diagram of the neural network low dimensional representation of the nodes, part of this document is to map neural network in smooth problem was analyzed in the process of information transfer. Then we analyzed the research methods and found that most of the literatures are about graph embedding algorithms. Some literatures study graph embedding algorithms based on probability graphs, and the other literatures study graph embedding algorithms based on deep learning In addition, the analysis of literatures shows that most literatures study semi-supervised and unsupervised graph neural networks, and some are fully supervised neural networks. Finally, we also analyze the measurement methods of various literatures, and find that some literatures study the accuracy of the proposed algorithm or model in node classification, and the other literatures analyze the efficiency of the proposed algorithm or model.

Through the analysis of the research methods and measurement methods of various literatures, the characteristics and innovation points of each article are found. Combined with the previous analysis, this paper also puts forward the future research direction of graph neural network and points that can be optimized.

The rest of the paper is organized as follows. Section II gives the classification of research objects of Unsupervised Graph Neural Network. Section III introduces the classification of research methods. Section IV introduces the comparison of experimental analysis in related literature. Section V discusses the research opportunities in future work and Section VI concludes the paper.

2 CLASSIFICATION OF RESEARCH OBJECTS

Table 1: Different Research Objects

Topics	Network classification		
	GCN	GNN	GAN
Low dimensional represent	I. [9][15][19][17]	II. [1][3]	III
Over-smoothing problem	IV. [8][13][7]	V.	VI.[16]

2.1 Criteria

Figure in recent years because neural network has attracted the attention of many scholars, the study of map neural network is also involved in many different neighborhood. In order to find out the differences, this paper will classify map neural network In addition, for neural network study there are two main problems, one is about the low dimensional representation of the nodes, the other is a smooth problem in the process of information transfer In this section, two separate and distinct criteria will be used to classify existing strategies into different types.

1) **Network classification**. There are three types: figure convolution network and the neural network and generate against network most of the existing map neural network is designed as a semi-supervised task, because the supervision information can reduce the messaging process of smoothing problem, and figure convolution network and the neural network can learn by graph embedding low-dimensional representation of the nodes.

2) **Topics**. There are two kinds of problems: Low-dimensional representation or over-smoothing of nodes. The purpose of graph embedding is to represent every node in the graph as a low-dimensional vector and to use that vector to perform a series of downstream tasks in semi-supervised learning tasks, using supervised information (labeled nodes) to mitigate over-smoothing of message delivery. Unfortunately, due to the lack of supervised information, the super-smooth problem tends to be more serious in unsupervised tasks. Therefore, the over-smooth problem of information transmission exists in

graph convolutional networks and generative adversative networks, while the low-dimensional representation problem of nodes only exists in graph convolutional networks and graph neural networks.

2.2 The Classification

Based on the appeal classification standard, we give the classification in Table 1. The meaning of each class is as follows:
Type I: This type is to study the low-dimensional representation of nodes in graph convolutional neural networks by learning the low-dimensional representation of nodes in complex networks through graph embedding algorithm.
Type II: This type is to study the low-dimensional representation of nodes in graph neural networks.
Type III: This type deals with low-dimensional representation of nodes in generative adversarial networks.
Type IV: This type studies the problem of over smooth information transfer in graph convolutional networks.
Type **V**：This type studies the over-smooth problem of information transfer in graph neural networks.
Type VI: This type deals with the over-smooth problem of information transfer in Graph generation network.

2.3 Explanation of Different Types

References ([9][15][19]) belong to Type I. Reference [9] studies how to effectively capture and measure the affinity relationship between nodes, and learn the low-dimensional representation of attributes and nodes in the same semantic space. Meanwhile, in order to obtain high-quality embedding, this paper proposes a variational auto-encoder, which emplaces the Gaussian distribution of the mean and Gaussian variance of each node and attribute. Reference[15] proposed a general method of learning node representation in graph structured data in an unsupervised way -- depth graph information analysis (DGI) DGI relies on the mutual information between the maximized patch representation and the corresponding high-level graph abstract Both are derived using the established graph convolutional network architecture. Reference [19] explores an architecture-jumping knowledge (JK) network that flexibly utilizes different neighborhood ranges for each node to achieve better structure-aware representation.
References ([1][3]) belong to Type II. Reference [1] studies learning low-dimensional vectors to represent vertices in graphs, and proposes a new GraRep model for learning weighted graph vertex representation. Different from existing work, this model integrates the global structure information of graphs into the learning process. Current feature learning methods are not capable of capturing the diversity of connection patterns observed in the network. Reference [3] Node2vec, an algorithm framework for learning continuous feature representation of nodes in the network, is proposed .In Node2vec, the mapping of nodes to low-dimensional feature Spaces is learned to maximize the possibility of preserving nodes' network neighborhoods.
References ([8][13][7])belong to Type IV . Reference [8] proved that graph convolution of GCN model is actually a special form of Laplacian smoothing form, which is the key reason why GCNs comes into play, but it also brings many potential problems of over-smoothing of convolution layer In order to overcome the limitations of GCN model with shallow architecture, this paper proposes collaborative training and self-training methods to train GCNs Reference [13] proposes a novel and flexible DropEdge technology to alleviate the over-smoothing and over-fitting problems DropEdge removes a random number of edges from the input graph at each training stage, like a data intensifier and a messaging decelerator. The paper also theoretically proves that DropEdge can either slow down over-smooth convergence or mitigate the resulting information loss. More importantly, DropEdge is a general skill that can be equipped with many other backbone models to improve performance. In reference [7], an improved propagation scheme based on personalized PageRank is derived using the relationship between graph convolutional network (GCN) and PageRank. This propagation process is used to construct a simple model, neural predictive personalized propagation (PPNP) and its fast approximation APPNP.

References ([16]) belong to Type VI. Reference [16] proposed a framework to solve the weakness of attentional restriction in graph attentional network and the problem of over-smoothing on decision boundaries. This paper first proved the over-smoothing behavior of GAT theoretically, and then developed a method using constrained attentional weights according to class boundaries and feature aggregation patterns In addition, in order to alleviate the problem of over-fitting, this paper proposes additional constraints on the graph structure.

3 CLASSIFICATION OF RESEARCH METHODS

Table 2. Different Research Methods

Graph embedding algorithm	Machine learning		
	Semi-supervised learning	Unsupervised learning	Supervised learning
Based on probability model	I. [15][10][11][17]	II. [3][20]	III.[1][14]
Based on deep learning	IV. [2]	V. [9][4][8][7][6]	VI

3.1 Criteria

Most existing graph neural networks (GNNS) are designed based on deep learning for semi-supervised learning tasks, in which supervised information (labeled nodes) is used to alleviate the over-smooth problem of message delivery. In this section, three independent and different criteria would be used to divide research objects into different types:

1) **Machine learning**. There are three types here: Semi-supervised learning, unsupervised learning or supervised learning. There are many methods of artificial intelligence, but the method used for graph neural network is machine learning, and most of the researches on graph neural network are semi-supervised learning, because supervised information can alleviate the over-smooth problem of information transmission. Therefore, three different machine learning methods will be studied in this paper.

2) **Graph embedding algorithm**. There are two kinds of Graph embedding algorithm here: Based on probability model or Based on deep learning. In various references, this paper uses two different graph embedding algorithms based on probability model and deep learning respectively to study graph clustering, so this is also an important distinguishing standard of this paper

3.2 The Classification

Based on the appeal classification standard, we give the classification in Table 2. The meaning of each class is as follows:

Type I: This type is to overcome the over-smooth problem of information transfer in semi-supervised graph neural network by adopting the embedded method based on probability graph model.

Type II: This type is to overcome the flatness of information transfer in unsupervised graph neural network by adopting the embedding method based on probability graph model.

Type III: This type overcomes the information transfer in fully supervised graph neural network by adopting the embedding method based on probability graph model.

Type IV: This type is to overcome the over-smooth problem of information transfer in semi-supervised graph neural network by adopting the embedding method based on deep learning model.

Type V: This type is to overcome the over-smooth problem of information transfer in unsupervised graph neural network by adopting the embedding method based on deep learning model.

Type VI: This type of embedding method based on probability graph model is used to overcome the over-smooth problem of information transfer in fully supervised graph neural network.

3.3 Explanation of Different Types

References ([15][10][11]) belong to Type I. Reference [15] proposed depth graph information (DGI), which is a general method to learn node representation in graph structure data in an unsupervised way. DGI relies on the mutual information between the maximized patch representation and the corresponding high-level graph summary, and these graphs are derived from the established graph convolutional network architecture The patch representation after learning summarizes the subgraph centered on the nodes of interest, so Reference can be reused in the downstream node intelligent learning task. In reference [10], DeepWalk is proposed to summarize recent advances in language modeling and unsupervised feature learning, or deep learning. From word sequences to graphics, DeepWalk uses local information from truncated random walks to learn potential representations by treating walks as equivalent sentences. Reference [11] uses graph Mutual information (GMI) to measure the correlation between input graph and high-level hidden representation. GMI extends the traditional calculation idea of mutual information from vector space to graph domain. It is essential to measure mutual information from two aspects of node characteristics and topological structure. The isomorphic transformation of the input graph is invariant, which is an unavoidable constraint in many existing graph representation learning algorithms. In addition, existing mutual information estimation methods, such as MINE, can be used to effectively estimate and maximize it.

References ([3][20]) belong to Type II. In Node2VEC, Reference [3] learns the method of mapping nodes to low-dimensional space of features, which can keep the network neighborhood of nodes to the maximum extent. The paper defines a flexible node network neighborhood concept and designs a biased random walk process, which can effectively explore different neighborhoods Reference [20] proposed a semi-supervised learning framework based on graph embedding. The paper developed two variants of the method, transduction and induction, by training the embedding of each instance to jointly predict the class tags and neighborhood context in the graph In the transformation variant of the method, class labels are determined by learned embeddings and input feature vectors, while in the inductive variant, embeddings are defined as parameter functions of feature vectors, so instances not seen in training can be predicted.

References ([1]) belong to Type III. Reference [1] proposed a new GraRep model for learning weighted graph vertex representation. This model represents the vertices in the graph by learning low-dimensional vectors. Different from existing studies, it integrates the global structure information of the graph into the learning process.

References ([12]) belong to Type IV. Reference [12] proposed a novel adversarial graph data embedding framework, which encodes the topological structure and node content of the graph into a compact representation, and on this basis trains the decoder to reconstruct the structure of the composition. In addition, the potential representation is matched with the prior distribution through the adversarial training scheme In order to learn a robust embedding, two adversarial methods are developed, adversarial regularized graph auto-encoder (ARGA) and adversarial variational graph auto-encoder (ARVGA).

References ([9][4][7]) belong to Type V. Reference [9] proposed an attribute network model (CAN) based on collaborative embedding, which CAN learn low-dimensional representations of attributes and nodes in the same semantic space, so as to effectively capture and measure the similarity between them In order to obtain high-quality embedding, we propose a variational auto-encoder Reference that uses the mean and variance of Gaussian distribution to embed each node .Reference [4] Introduced in many network is another feature, namely the characteristics of community structure in this feature, the network node to close together, and only a loose connection between these groups We proposed a method to detect such community, based on using the centricity index to find the idea of the community boundaries Reference [7] Using the

relationship between graph convolution network (GCN) and PageRank, we derive an improved propagation scheme based on personalized PageRank. We use this propagation process to construct a simple model, personalized neural predictive propagation (PPNP), and its fast approximation, APPNP.

4 REVIEW OF EXPERIMENTAL ANALYSIS

In this section, we will classify the metric of evaluation and system parameters, as shown in Table 3. In Table 3, all experimental analysis is also classified according to the metric and parameters. It can be seen from Table 3 that most of the references compare superiority, effectiveness index of algorithm.

Table 3. Experiments with Different Metric and Parameters

Metric	Parameters	
	algorithm	model
Accuracy	I. [15][8][6][13][11][16]	II. [18][7][20][17]
Efficiency	III. [2]	IV. [1][14]
Others	V. [3]	VI.

4.1 Metric of Evaluation

Accuracy means the proportion of correct node classification to actual node classification of social network. The formula is as follows:

$$\text{Accuracy} = \frac{\text{Number of correctly categorized node}}{\text{Number of actually categorized node}}$$

Efficiency means that the time complexity required to use the algorithm or model is less than that of the existing algorithm or model. The formula is as follows:

$$T(n) = O(f(n))$$

Other metric includes Macro-F1 Score, Micro-F1 Score and Normalized mutual information (NMI).

4.2 System Parameters

The algorithm represents the algorithm used for node classification in social network. The time complexity of the algorithm is small, and the division of nodes in social network is more accurate than the existing algorithm. However, the information in the semi-supervised graph neural network can alleviate this problem.

Model represents the model used for node classification. When dealing with different problems, different models can achieve different effects. There are also some other parameters, such as different experiments of graph neural network and comparison experiments of existing models.

4.3 Experimental Comparison

In reference [15], the authors tested the proposed method on a synthetic network. For the transformation task, the average classification accuracy (with standard deviation) at the test nodes of the proposed method was reported after 50 training sessions (followed by logistic regression), indicating the high accuracy of the DGI method.

In reference [8], the author conducted a large number of experiments on real benchmark to verify the proposed theory and method, including joint training self-training combination and cross experiment results show that the proposed joint training and self-training methods have high accuracy in node classification of GCNs.

In reference [6], the authors evaluate the performance of the proposed method BTLSC detection community in nine real world networks and real communities on the ground. Experimental results show that the proposed method has high accuracy in node classification

In reference [13], the authors tested the layer 2, 8, 32 of the proposed model on different data sets, and the results show that the proposed model DropEdge improves the accuracy of the test.

In reference [11], the author compares the proposed method with other unsupervised methods. Experimental results show that the proposed GMI-Mean and GMI-Adaptive methods achieve the best classification accuracy on all three data sets. This powerful performance benefits from the encoded representation that preserves the node features and topology information in the graph to the maximum extent, which is conducive to classification。

In reference [1], experiments are carried out on three different types of tasks, including clustering classification and visualization experiments. The results show that the proposed GraRep model can integrate different k-step local relation information into the global graph representation of different types of graphs, so it can be effectively applied to different tasks.

In reference [14], the author compares the proposed LINE model with several existing graph embedding methods that can be extended to very large networks. The results show that in the reconstructed network, LINE (second) is superior to DeepWalk's combination of first and second order proximity on the original network in most cases and has captured most of the information, LINE. The method is a very effective and efficient network embedding method, which is suitable for dense and sparse networks.

In reference [7], the author conducted experiments on GCN, N-GCN, GAT, BootStrapped feature Propagation and JK models respectively. According to the experimental results, the proposed model personalized Propagation of Neural Predictions (PPNP) has high accuracy in node classification.

In reference [20], compared with the existing methods, the author shows that the proposed semi-supervised learning framework based on graph embedding has high accuracy in text classification

In reference [17], the author evaluated the node classification performance of the proposed MODEL SGC on Cora, Citeseer and Pubmed datasets respectively. Experimental results show that the SGC has better performance on Citeseer, which is due to fewer parameters of SGC and therefore less over-fitting.

In reference [2], the performance of dual-Sinkhorn is tested on MNST dataset, and the experimental results show that the proposed algorithm is efficient and has low time complexity.

5 DISCUSSION AND SUGGESTION

This paper discusses the research method and research object found figure of the neural network study mostly are designed for a semi-supervised task, because supervision information can slow the smoothing problem of messaging But smooth information transmission problem in unsupervised way more serious, because there is no degree information is available Therefore, unsupervised neural network research is an important issue Deep learning undoubtedly become the preferred method However, to date, most research on and the figure of the neural network are based on the graph embedding algorithm without supervision and supervision, and supervision of all research methods based on depth of learning algorithm were rare .Therefore, this article puts forward the following several direction, can offer the direction for the future study of the figure of neural network.

1) The low-dimensional representation of nodes in graph attention network can be studied, and then the graph embedding algorithm based on deep learning is used for analysis. Finally, the measurement is carried out by the accuracy of node classification.

2) Many practical methods are selected in this paper and compared with other methods Therefore, future research can conduct in-depth research on the accuracy and efficiency of the method and find the best method for different objects

6 CONCLUSIONS

Through the previous analysis, we find that most of the researches on graph neural networks are based on semi-supervised learning tasks, but the over-smooth problem of message delivery is more serious in unsupervised learning Therefore, the graph neural network based on unsupervised mode can be further studied in the future, especially the graph embedding algorithm based on deep learning to learn the low-dimensional representation of nodes in graph attention network. There are few analyses and researches in this area. Therefore, this kind of research method can be further studied in the future.

References

[1]S. Cao, W. Lu, and Q. Xu, "Grarep: Learning graph representations with global structural information," in CIKM, 2015, pp. 891–900.

[2] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," in NIPS, 2013, pp. 2292–2300.

[3]A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in SIGKDD, 2016, pp. 855–864.

[4] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," PNAS, vol. 99, no. 12, pp. 7821–7826, 2002.

[5]D. Jin, K. Wang, G. Zhang, P. Jiao, D. He, F. FogelmanSoulie, and X. Huang, "Detecting communities with multiplex semantics by distinguishing background, general and specialized topics," 2019, pp. 1–1.

[6]T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in ICLR, 2017.

[7]J. Klicpera, A. Bojchevski, and S. Gunnemann, "Predict then propagate: Graph neural networks meet personalized pagerank," in ICLR, 2019.

[8]Q. Li, Z. Han, and X. Wu, "Deeper insights into graph convolutional networks for semi-supervised learning," in AAAI, 2018, pp. 3538–3545.

[9]Z. Meng, S. Liang, H. Bao, and X. Zhang, "Co-embedding attributed networks," in WSDM, 2019, pp. 393–401.

[10]B. Perozzi, R. Al-Rfou, and S. Skiena, "Deepwalk: online learning of social representations," in SIGKDD, 2014, pp. 701–710.

[11]Z. Peng, W. Huang, M. Luo, Q. Zheng, Y. Rong, T. Xu, and J. Huang, "Graph representation learning via graphical mutual information maximization," in WWW, 2020, pp. 259–270.

[12]S. Pan, R. Hu, G. Long, J. Jiang, L. Yao, and C. Zhang, "Adversarially regularized graph autoencoder for graph embedding," in IJCAI, 2018, pp. 2609–2615.

[13]Y. Rong, W. Huang, T. Xu, and J. Huang, "Dropedge: Towards deep graph convolutional networks on node classification," in ICLR, 2020.

[14]J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei, "LINE: large-scale information network embedding," in WWW, 2015, pp. 1067–1077.

[15]P. Velickovic, W. Fedus, W. L. Hamilton, P. Lio, Y. Bengio, and R. D. Hjelm, "Deep graph infomax," in ICLR, 2019.

[16]G. Wang, R. Ying, J. Huang, and J. Leskovec, "Improving graph attention networks with large margin-based constraints," NerIPS Workshop, vol. abs/1910.11945, 2019.

[17]F. Wu, A. H. S. Jr., T. Zhang, C. Fifty, T. Yu, and K. Q. Weinberger, "Simplifying graph convolutional networks," in ICML, 2019, pp. 6861–6871.

[18]K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?" in ICLR, 2019.

[19]K. Xu, C. Li, Y. Tian, T. Sonobe, K. Kawarabayashi, and S. Jegelka, "Representation learning on graphs with jumping knowledge networks," in ICML, 2018, pp. 5449–5458.

[20]Z. Yang, W. W. Cohen, and R. Salakhutdinov, "Revisiting semi-supervised learning with graph embeddings," in ICML, 2016, pp. 40–48.